TABLE 1

Length of Internal Medicine Clerkships as Related to Numbers of Patient Work-Ups per Student and Numbers of Disease Categories Seen by 143 Students at 42 U.S. and Canadian Medical Schools, 1979–1982

Clerkship Length in Weeks	No. of Schools	No. of Students	No. of Patient Work-Ups Per Student				No. of Disease Systems Seen		
			Maximum	Minimum	Median	Per Week	Maximum	Minimum	Mean
12	20	107	33	6	19	1.6	15	2	9.9
8	8	20	32	8	14	1.9	13	5	9.1
6	5	16	21	7	11	2	10	4	7.3

ber of students in the clinical rotation groups, ranging from one to 16. Because this questionnaire item was not more specific, the authors were unable to determine how many students competed for the available patients in each clinical setting.

To study patient-related experience by disease system, systemic categories of disease were selected from standard textbooks of internal medicine, and the experiences of the 165 students who responded to questions on their patients were sorted into these categories. The students saw a total of 2,891 patients. Because diseases seldom fit into clear-cut categories and many diseases cross several systems, it was important to make consistent judgments as to the major disease system involved in categorizing each patient. For example, patients with diabetes mellitus were placed in the metabolic category, and those with alcoholic cirrhosis were listed under hepatobiliary disease rather than nutritional disease, which might also have been appropriate.

Table 2 gives the number and percentage of patients in each disease category. A majority of the students' clinical experiences were obtained from patients with heart disease, cancer, and respiratory disorders. Very few patients were seen with primarily dermatologic, critical care or accidents, and neuromuscular disorders.

To gauge the breadth and appropriateness of students' experiences with patients, the authors examined the experiences of students who saw a small number of patients, those who saw a large number of patients, and those who, in the authors' opinion, may have had unbalanced experiences. Table 3 displays the low, median, and high levels of experiences with

TABLE 2

Numbers and Percentages of 2,891 Patients in 21 Internal Medicine Disease Categories Seen by 165 Students at 42 U.S. and Canadian Medical Schools, 1979–1982

Disease Category	No. of Patients	Percentage of Total
Heart	422	14
Oncology	402	14
Respiratory	313	11
Alimentary	228	8
Metabolic, diabetes mellitus	212	7
Infectious disease	207	7
Central nervous system	166	6
Vascular	158	5
Genitourinary	119	4
Hepatobiliary, alcoholic cirrhosis	116	4
Drug intoxication and abuse	102	4
Hematology	82	3
Immune	70	2
Bone and joint	51	2
Endocrine	51	2
Pancreas	50	2
Behavioral/psychiatric	49	2
Neuromuscular	35	1
Critical care/accident	28	*
Dermatology	20	*
Nutritional	10	*

* Less than 1 percent.

TABLE 3

Numbers of Patients, Numbers of Disease Categories, and Percentages of Patients in One Disease Category Seen By 22 Selected Students in Required Internal Medicine Clerkships, By Levels of Patient-Related Experience, in 16 Medical Schools, 1979–1982

	12-week Clerkship			8-week Clerkship			6-week Clerkship			Unbalanced Clerkships (12 Week)		
Levels of Experience*	No. of Pat. Seen	No. of Dis. Categ. Seen	Perc. of Pat. in One Dis. Categ. Seen	No. of Pat. Seen	No. of Dis. Categ. Seen	Perc. of Pat. in One Dis. Categ. Seen	No. of Pat. Seen	No. of Dis. Categ. Seen	Perc. of Pat. in One Dis. Categ. Seen	No. of Pat. Seen	No. of Dis. Categ. Seen	Perc. of Pat. in One Dis. Categ. Seen
Low												
Student 1	6	6	17	9	6	50	7	5	40	16	7	56
Student 2	6	4	33	8	7	25	6	4	67	21	8	62
Median												
Student 3	19	7	42	14	10	28	11	7	36	10	2	90
Student 4	19	10	26	14	5	36	11	8	27	10†	4	60
High												
Student 5	35	13	28	32	13	12	21	9	28			
Student 6	33	14	18	20	13	20	21	10	33			

* Levels of experience indicate those students who obtained histories and performed physical examinations on the fewest number of patients, on the most number, and on a median number of patients in each clerkship category.

† Clerkship length was 6 weeks.

patients for 22 selected students in clerkships of 12, 8, and 6 weeks' duration. For each clerkship group displayed, the authors selected two of the students who saw the fewest patients, two who saw the median number, and two who saw the most patients. The assumption is that students who see too few patients and those who see too many may not have optimum experiences. Listed for each selected student are the total number of patients seen, the total number of disease systems into which his patients could be placed, and the percentage of patients who fell into only one disease category. The experiences of these selected students emphasize the variability within clerkships of the clinical experience in terms of the number of patients and disease systems with which the student comes in contact. A vast majority of the 22 students had patient-related experiences within several disease systems. However, a few recorded experiences which might be considered unbalanced, since 56, 60, 62, 67, and 90 percent of their patients fell within one category of disease.

A total of 176 students designated those physicians who supervised their clinical experience, heard presentations, checked write-ups, and offered constructive criticism. These teaching and supervisory responsibilities were almost always shared by faculty members, residents, and interns, with the combination of faculty members and residents shouldering the major responsibility. The faculty assumed major roles in supervision (99 percent) and hearing the findings of student workups (96 percent). Faculty members were less involved in checking write-ups (59 percent) and offering constructive criticism (66 percent). Students cited residents as giving supervision (88 percent), hearing presentations of history and physical examinations (69 percent), checking write-ups (59 percent), and giving constructive criticism (57 percent). Interns were least

involved in clerkship responsibilities, although 40 percent of the students cited them as sharing supervision. It should be noted that all students had the opportunity to present their patient work-ups. Only one of the 176 students reported that her write-ups were not checked, and only two students said they received no constructive criticism.

Discussion

To study the qualitative and quantitative aspects of the required clerkship in internal medicine, the authors analyzed the patient-related experience of 180 medical students at 42 U.S. and Canadian medical schools who took histories and performed physical examinations on 2,891 patients. The length of the clerkships varied from six to 15 weeks. Similar variability was found in the number of hours worked per day and per week, the number of patients seen, and the amount of supervision by faculty members and residents. Unsurprisingly, the authors found that the students reported working long hours (an average of 10.4 hours per day and 67.8 hours per week).

The students saw a variety of disease processes, and these related in a general way to the length of their clerkship experiences. The most frequent categories of disease seen by the students were heart disorders, cancer, and disease of the respiratory system. Much less experience was recorded in dermatologic, critical care and accident problems, and neuromuscular disorders. The lack of student experiences in dermatology, critical care, accident, and neuromuscular disorders during the required clerkship reflects more the specialization of medical centers as providers of tertiary care than a deficiency in general medical teaching. It should be emphasized that a vast majority of the nutritional problems in the study were related to alcoholic liver disease and were tabulated under the hepatobiliary heading. The small number of cases in skin, critical care and accident, neuromuscular, and behavioral/psychiatric headings is not necessarily a cause for alarm, since a majority of medical students have clerkship experiences in surgery and psychiatry; critical care, neurology, and dermatology are offered as fourth-year elective experiences. The students' patient-related experiences compared favorably with mortality rates in the United States in 1981, when the leading causes of death were diseases of the cardiovascular system, including heart, atherosclerosis, and hypertension; malignant neoplasms; cerebrovascular diseases, accidents, poisoning, and violence; diseases of the respiratory system; and diseases of the digestive system (3).

Students' experiences with patients were examined in terms of the number of patients seen. The question of breadth of experience may be raised for students who saw few patients. In the present survey, the authors did not evaluate the quality of a student's experience in terms of an in-depth study or thorough knowledge on the part of the student of the individual patient. For example, a student who prepared a thorough work-up and did readings on six patients could have had a superior educational experience to that obtained if he had worked on 20 to 30 patients whose disease processes were neither carefully studied nor understood. The authors are concerned about the students who seemed to have unbalanced experiences (Table 3). Undoubtedly, a great deal may be learned about a group of patients within one category of disease, but seeing 56 to 90 percent of patients in one disease category may not provide appropriate experience for a medical student in a required clerkship in internal

medicine. The critical issue here is assignment and monitoring of a student's clinical experience, and it is possible that in some schools this issue needs attention. Perhaps a responsible teacher should monitor an individual student's experience to ensure that it meets the objectives of the clerkship. Asking students to maintain a log of their patient-related experiences might prove helpful.

In summary, the preliminary survey reported here has provided some quantitative measures of internal medicine clerkships in terms of hours, patients seen, and supervision, but it does not answer the basic question of whether medical schools are neglecting the teaching of clinical skills and supervision of clinical training. Because of the great variability in numbers of patients seen and the indication that some students may see too few different disease categories, further questions should be asked. How does the medical student spend his time during the clerkship? Who observes the student who takes a history and performs a physical examination? What rounds does the student attend and/or participate in, and what responsibilities does he accept? Is the student treated as an observer or an active participant? Such information in quantitative terms is necessary for the proper evaluation of medical students' basic clinical experience.

References

1. ENGEL, G. L. Are Medical Schools Neglecting Clinical Skills? *J.A.M.A.*, **236**: 861–863, 1976.
2. WIENER, S., and NATHANSON, M. Physical Examination: Frequently Observed Errors. *J.A.M.A.*, **236**:852–855, 1976.
3. Mortality from Selected Courses, United States, 1971 and 1981. In *Basic Data Relating to the National Institutes of Health*. Washington, D.C.: U.S. Department of Health and Human Services, 1985, P. 57.

Vol. 61 No. 4 April 1986

JOURNAL OF Medical Education

PUBLICATION OF THE ASSOCIATION OF AMERICAN MEDICAL COLLEGES

The Effect of Commercial Coaching Courses on Performance on the MCAT
Robert F. Jones

Race and Socioeconomic Status in Medical School Choice and Indebtedness
Gloria J. Bazzoli, Kathleen Adams, and Sara L. Thran

Evaluation of Ambulatory Care Training by Graduates of Internal Medicine Residencies
Lawrence S. Linn, Robert H. Brook, Virginia A. Clark, Arlene Fink, and Jacqueline Kosecoff

Charges by Residents and Faculty Physicians in a University Hospital Pediatric Practice
Jay E. Berkelhamer

Students' Perceptions of Values Emphasized in Three Medical Schools
Brigitte Maheux and François Béland

Supplement
Commentary on the GPEP Report

Journal of Medical Education

Volume 61 • Number 4 • April 1986

The Effect of Commercial Coaching Courses on Performance on the MCAT—Robert F. Jones 273

Race and Socioeconomic Status in Medical School Choice and Indebtedness—Gloria J. Bazzoli, Kathleen Adams, and Sara L. Thran 285

Evaluation of Ambulatory Care Training by Graduates of Internal Medicine Residencies—Lawrence S. Linn, Robert H. Brook, Virginia A. Clark, Arlene Fink, and Jacqueline Kosecoff 293

Charges by Residents and Faculty Physicians in a University Hospital Pediatric Practice—Jay E. Berkelhamer 303

Students' Perceptions of Values Emphasized in Three Medical Schools—Brigitte Maheux and François Béland 308

Communications: A Case Study of a Medical School Special Admissions Program—Percy J. Russell, Nolan E. Penn, Harold Simon, Teresa Jacob, Eddie Castro, and Claire Stafford 317

Comparison of the Abbreviated and Original Versions of the Myers-Briggs Type Indicator Personality Inventory—Lisa I. Leiden, Tracy L. Veach, and Michael W. Herring 319

Preventive Health Measures Carried Out by Residents During New and Follow-Up Patient Visits—Carol Cheney and Joe W. Ramsdell 321

Performances of Female Medical Students in an Obstetrics and Gynecology Clerkship—Warren C. Plauché and Joseph M. Miller, Jr. 323

Using Filing Cards To Document and Computerize Procedures Performed by Internal Medicine Residents—Kurt Kroenke and John L. Carpenter 326

Teaching Critical Appraisal and Application of Medical Literature to Clinical Problem-Solving—Kenneth L. Radack and Barbara Valanis 329

Briefs 331

Editorial: Should Higher Diagnostic and Therapeutic Costs Be Justified for Educational Reasons Alone? 333

Datagram: A Comparison of the Predictive Validity of the MCAT for Coached and Uncoached Students 335

Bibliography 339

Supplement: Commentary on the Report of the Panel on the General Professional Education of the Physician and College Preparation for Medicine 345

Officers of the Association and Members of the Executive Council, 1985–86

CHAIRMAN: Virginia V. Weldon, *Washington University*
CHAIRMAN-ELECT: Edward J. Stemmler, *University of Pennsylvania*
IMMEDIATE PAST CHAIRMAN: Richard Janeway, *Wake Forest University*
PRESIDENT: John A. D. Cooper, *AAMC, Washington, D.C.*

COUNCIL OF DEANS
D. Kay Clawson, *University of Kansas**
Arnold L. Brown, *University of Wisconsin*
William T. Butler, *Baylor College of Medicine*
Robert S. Daniels, *University of Cincinnati*
William B. Deal, *University of Florida*
Louis J. Kettel, *University of Arizona*
Richard H. Moy, *Southern Illinois University*
John Naughton, *SUNY at Buffalo*
Richard S. Ross, *Johns Hopkins University*

COUNCIL OF ACADEMIC SOCIETIES
David H. Cohen, *SUNY at Stony Brook**
William F. Ganong, *UC, San Francisco*
Frank G. Moody, *University of Texas, Houston*
Virginia V. Weldon, *Washington University*

COUNCIL OF TEACHING HOSPITALS
C. Thomas Smith, *Yale-New Haven Hospital**
J. Robert Buchanan, *Massachusetts General Hospital*
Spencer Foreman, *Sinai Hospital of Baltimore*
Sheldon S. King, *Stanford University Hospital*

ORGANIZATION OF STUDENT REPRESENTATIVES
Richard Peters, UC, San Diego*
Vicki Darrow, *University of Washington*

DISTINGUISHED SERVICE MEMBER
Charles C. Sprague, *University of Texas, Southwestern Medical School*

* Chairman of respective organization.

The Effect of Commercial Coaching Courses On Performance on the MCAT

Robert F. Jones, Ph.D.

Abstract—Two studies on the relationship between taking a commercial coaching course and performance on the Medical College Admission Test (MCAT) are reported. In both studies, one based on repeating examinees and another based on first-time examinees, differences in scores favoring coached examinees were found on the Science Knowledge and Science Problems subtests. The results on the Skills Analysis: Quantitative subtest were mixed, while neither study found differences between the two groups in scores on the Skills Analysis: Reading subtest. The analyses further revealed that the differences in scores on the Science Knowledge and Science Problems subtests were less pronounced for examinees with low scores on the Skills Analysis: Reading and Skills Analysis: Quantitative subtests, examinees with very low or very high undergraduate grade-point averages, and examinees from very selective undergraduate colleges. The substantive importance of the differences was also explored. The coaching effect was estimated to increase an average student's probability of acceptance by 5 percentage points. The results suggest that coaching for the MCAT has a limited effect that supports rather than undermines the validity of the test.

A growing debate has emerged in recent years concerning the effects that commercial test-preparation services or coaching courses have on students' performance on standardized tests. The debate was joined in 1979 with the publication of two reports (1, 2) by the Federal Trade Commission (FTC), which had launched an investigation of commercial services preparing students to take the Scholastic Aptitude Test (SAT) and the Law School Admission Test (LSAT). The two reports differed in their conclusions regarding the effectiveness of coaching for these tests. Much of the research literature on coaching, before and after the FTC reports, has tended to focus on the SAT. Messick (3) provided a reanalysis of the FTC data from several perspectives, and Messick and Jungeblut (4) wrote a review of studies of coaching for the SAT. They concluded that short-term coaching for the SAT generally has only minor effects on performance, although countervailing views persist (5).

The coaching issue has more recently surfaced in research in medical education. Scott and colleagues (6) reported significantly higher scores on Part I of the National Board of Medical Examiners (NBME) examinations for a group of stu-

Dr. Jones is director of institutional studies, Department of Institutional Development, Association of American Medical Colleges (AAMC), Washington, D.C. He formerly was projects director, Medical College Admission Test technical and research activities, Division of Educational Measurement and Research.

For related datagram, see page 335.

dents enrolled in a commercial coaching course compared with a control group matched on basic science grade-point average. Scott's findings contrasted with the results of a similar study by Lewis and Kuske (7), who failed to find any differences in performance on an NBME examination between coached and uncoached students. Anderson (8) raised the coaching issue with regard to the Medical College Admission Test (MCAT), but no studies have yet been reported.

The data reported here present the results of two studies designed to determine differences associated with taking commercial coaching courses on the performance of first-time and repeating MCAT examinees.

Background

Resolution of the controversy over the effects of coaching on performance has generally been hampered by differences in the meaning given to the term coaching and misconceptions about the nature of aptitude and achievement that standardized tests purport to measure (9). Cole (10) lists six possible components of test-preparation programs, each of which has varying implications for test validity: supplying correct answers (cheating), taking the test for practice, maximizing motivation, optimizing test anxiety, instructing in test-taking skills, and instructing in test content. The first component is quite obviously associated with the negative connotations of coaching and poses a definite threat to test validity. Instruction in test-taking skills, the fifth component, may also have a negative connotation if the test-taking instruction is specifically directed to flaws in test construction that cue students to the correct answer. As with the first component, students coached in this manner may achieve scores that overestimate their knowledge and skills. Instruction in general test-taking skills (for example, following instructions carefully and allotting time properly), and in methods for increasing motivation and controlling anxiety and practice in test-taking do not enable students to achieve scores that overestimate their true level of knowledge and skills. Rather, such instruction reduces the chances of underperforming. However, insofar as these factors account for differences in scores, students coached in this manner may have an advantage that lowers the overall validity of the test. The last component, instruction in test content, departs radically from the popularly held connotations of coaching but is an important element of many programs. The objective of such instruction is to increase the test-taker's knowledge and skill and as such does not necessarily lower test validity.

The relative emphasis given to each of these components within a coaching course tends to differ, depending upon whether the test is measuring aptitude or achievement. Aptitude tests like the SAT and LSAT assess basic analytical skills that are developed formally and informally over a long period of time. The underlying assumption that has driven researchers to focus on differences in performances resulting from coaching for these tests is that the abilities being tested should be relatively resistant to short-term intervention. If significant and substantial effects of coaching on performance were found for this kind of test, which includes the MCAT Skills Analysis (SA): Reading and Skills Analysis: Quantitative subtests, the validity of the test may be reduced, since the difference in the examinees' performances may derive from one of the components of coaching programs other than instruction in content that Cole (10) identified.

Achievement tests on the other hand, like the NBME examinations or the MCAT Science Knowledge and Science Problems subtests, by their nature reflect the influence of instruction and of study and review of the content areas covered. Coaching for achievement tests generally includes a structured review of content areas in addition to the usual advice on test-taking strategy and practice. To the extent that students who enroll in test-preparation courses for NBME examinations or the MCAT Science Knowledge and Science Problems subtests engage in a formal review of the subject matter covered, they might plausibly expect some enhancement of performance. In fact, the absence of any effect on these tests from well organized study and review over time, whether conducted in concert with a formal course or independently, would bring into question the construct validity of these tests. The effect of commercial coaching courses on performance on the NBME examinations or the MCAT Science Knowledge and Science Problems subtests threatens the validity of these tests only to the degree to which the magnitude of the effect exceeds reasonable expectations based on the nature and duration of the review program or to the extent to which the effect changes the predictive relationship between the test-taker's performance and subsequent academic performance.

Coaching and Test Repeaters

In the first study, the effects of commercial coaching courses on performance on the MCAT were examined using five samples of test repeaters from the years 1977 through 1981.

METHOD

Data were obtained from the MCAT program files of the Association of American Medical Colleges (AAMC). In each year, those examinees were identified who took the MCAT for the first time at the spring administration and then repeated the test the following fall. Examinees who had taken a commercial coaching course prior to the spring administration or those for whom that information was not available were eliminated from each sample. The remaining examinees were divided into those who took a commercial coaching course prior to the fall administration and those who did not. Information on the examinees' coaching course experience was based on their self-report on a questionnaire given at each test administration. The final samples for each of the years 1977 through 1981 were 5,014, 4,132, 3,318, 3,087, and 3,047 test repeaters, respectively. The percentages of repeaters in each sample who had taken a commercial coaching course between test administrations were 15, 19, 21, 26, and 32, respectively.

The study design enabled the scores of each group on the fall test administration to be compared, with each group's scores from the spring test administration serving as the baseline measure. It is important to note that the study design was not a true experimental design. The examinees were neither randomly selected nor randomly assigned to groups. The myriad other factors that might differentiate the two groups in addition to their coaching course experience and that might be related to gains in scores could not be controlled. Among the most important of these possible factors was the examinees' motivation to perform well, as reflected in the number of hours spent in preparation for the test. Therefore, any differences found in the scores of the two groups could be interpreted only as being associated with taking a commercial coaching course and not necessarily as

being due to the course itself. Despite this limitation, the importance of the issues surrounding commercial coaching courses more than justified pursuing the analyses.

The study design lent itself to a straightforward comparison of gains in scores as a way of determining differences associated with taking a review course, although certain statistical adjustments had to be made. Simple comparisons of gains have been shown to be misleading when the groups differed in their initial performance (11). In general, individuals with lower initial performance tend to make greater gains on retesting, all other things being equal, because of certain measurement properties of the scores. In the present study, analysis of covariance was used to adjust the mean scores of the fall group for differences in initial test performances. From these data, new adjusted gains for each group were derived by computing the difference between each group's adjusted mean performance on the fall administration and the mean performance for both groups of examinees combined on the spring administration.

Further analyses were undertaken to determine whether any performance differences found between coached and uncoached examinees varied among specific subgroups. For example, did women receive the same benefits from coaching courses as men did, or minorities the same as nonminorities? The subgroup characteristics selected for examination were sex; age (over 22 and 22 or younger); minority status (blacks, American Indians, Mexican-Americans, mainland Puerto Rican, and all others), undergraduate major (biological science, premedical science, physical science, mathematics and/or statistics, social science, humanities, and all others); level of course work in biology, chemistry, and physics; and level of performance on the SA: Reading and SA: Quantitative subtests. Statistical tests to measure any significant interaction effect in analysis of covariance were performed with one of the study samples.

RESULTS

Table 1 compares the average gains in scores, adjusted for initial score differences, of coached and uncoached examinees. These gains should be viewed in the context of the MCAT score scale, which ranges from 1 to 15 (15 is the highest score). The mean of the norm group (spring 1977 examinees) was 8.0 and the standard deviation 2.5 in each area tested. Comparisons are made in Table 1 for each area of assessment across the five samples.*

In 28 of the 30 comparisons (five samples times six subtests), a difference in gains was found that favored those taking commercial coaching courses. In 24 of these 28, the differences were statistically significant ($p < .01$). Differences were observed at three levels of magnitude. Scores on the Science Knowledge and Science Problems subtests showed differences that were similar to each other and were the largest overall. The coached examinees tended to improve their scores approximately one to one-and-a-half scale score points on the repeat administration; uncoached examinees improved their scores approximately one-half to one scale score point. The differences between the groups were approximately one-half of a scale score point. Evidence of a trend in these differences over time was mixed. Scores in the Biology, Physics, and Science Problems areas showed a generally upward trend, but the major increase on these subtests was between

* A detailed summary of these and other data reported in this study is available from the author.

TABLE 1

Adjusted Mean Gains in Scores of 18,598 Coached and Uncoached Examinees Who Repeated the Medical College Admission Test (MCAT), 1977 through 1981*

MCAT Subtests/Examinees	Year†					Average Across Years
	1977	1978	1979	1980	1981	
Biology						
Coached examinees	1.15	1.44	1.44	1.19	1.04	1.25
Uncoached examinees	.77	.99	.88	.58	.42	.73
Difference in gains	.38	.45	.56	.61	.62	.52
Chemistry						
Coached examinees	1.23	1.11	1.24	1.07	1.01	1.13
Uncoached examinees	.72	.64	.87	.69	.52	.69
Difference in gains	.51	.47	.37	.38	.49	.44
Physics						
Coached examinees	1.16	1.47	1.12	1.13	1.20	1.22
Uncoached examinees	.89	.94	.49	.56	.49	.67
Difference in gains	.27	.53	.63	.57	.71	.55
Science Problems						
Coached examinees	1.12	1.48	1.26	1.16	1.14	1.23
Uncoached examinees	.88	1.08	.74	.66	.55	.78
Difference in gains	.24	.40	.52	.50	.59	.45
SA: Reading						
Coached examinees	.60	.90	.52	.48	.84	.67
Uncoached examinees	.66	.79	.43	.49	.75	.62
Difference in gains	−.06‡	.11‡	.09‡	−.01‡	.09‡	.05‡
SA: Quantitative						
Coached examinees	.98	.96	.91	.61	1.02	.90
Uncoached examinees	.81	.77	.58	.30	.77	.65
Difference in gains	.17‡	.19	.33	.31	.25	.25

* For explanation of adjusted scores, see text.

† Sample sizes: 1977 = 759 coached, 4,255 uncoached; 1978 = 799 coached, 3,333 uncoached; 1979 = 710 coached, 2,608 uncoached; 1980 = 817 coached, 2,270 uncoached; 1981 = 989 coached, 2,058 uncoached.

‡ $p > .01$; all other differences are statistically significant.

1977, the first year of the new MCAT, and 1978. Scores on the Chemistry subtest, showing the largest differences in 1977, was the exception to this trend.

The data for SA: Quantitative subtest scores reflected a smaller difference. Both the overall gains in scores and the differences between the coached and uncoached groups were lower than those observed on the Science Knowledge and Science Problems subtests. The advantage for those taking commercial coaching courses was reduced to one-fourth of a scale score point when averaged across samples. This difference was statistically significant in four of the five samples.

The smallest gains were observed for the SA: Reading subtest. Differences in gains between the two groups were indistinguishable, consistently failing to achieve statistical significance in all five samples studied.

The effects of coaching courses were not found to differ significantly between examinee subgroups, except those based on the level of the examinees' initial score on the SA: Reading and SA: Quantitative subtests. The examinees with lower scores

on these subtests, and presumably weaker skills, tended to benefit less from the coaching course in terms of improving their performance on the science areas of assessment than examinees with average or high scores. Minority examinees in each sample of test repeaters tended to cluster among the examinees with the lower scores on the SA: Reading and SA: Quantitative subtests. This confounding made it impossible to determine whether minority examinees as such benefited less from commercial coaching courses than did nonminority examinees.

Figure 1 illustrates how the effects of coaching courses on performance were related to the test-taker's reading and quantitative skills. Three groups of examinees were defined by the average of their initial scores on the SA: Reading and SA: Quantitative subtests: low-skills score (mean = 1–6.5); medium-skills score (mean = 7.0–8.0); high-skills score (mean = 8.5–15). Within each of these groups, the differences in gains in scores on the Science Knowledge and Science Problems subtests between coached and uncoached examinees were computed. For economy of presentation, the differences were averaged across Science Knowledge and Science Problems subtest scores and were plotted in Figure 1 for each of the five samples. It should be noted that this averaging gives a double weight to Science Problems scores, as explained in the *MCAT Interpretive Manual* (12).

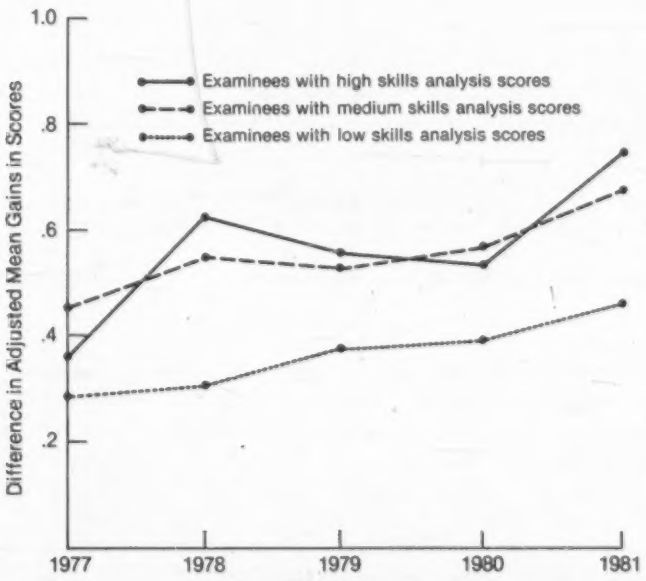

FIGURE 1

Differences in adjusted mean gains in scores of 4,074 coached repeat examinees on Science Knowledge and Science Problems subtests of the Medical College Admission Test (MCAT), compared with 14,524 uncoached repeat examinees, by level of examinees' scores on the Skill Analysis: Reading and Skills Analysis: Quantitative subtests of the MCAT, 1977 through 1981 (for explanation of adjusted scores and the three skills analysis groups, see text).

Coached examinees in each skills analysis group tended to outperform uncoached examinees, as indicated by a positive difference across groups and samples. The differences were consistently smaller for the low-skills score group than for the medium- and high-skills score group.

Coaching and First-Time Examinees

In the second study, the effect on performance of taking a commercial coaching course was investigated using data on examinees taking the test for the first time.

METHOD

Since examinees themselves choose whether or not to participate in commercial coaching courses, this study required finding a way to control for prior differences in ability and achievement between those who had enrolled in the course and those who had not. Two measures were used: undergraduate GPA and "selectivity" of undergraduate college. Both have been shown to correlate strongly with MCAT scores, and their use together provided an estimate of the examinee's expected level of performance on the MCAT (13).

The population consisted of all 1981–82 medical school applicants ($n = 22{,}618$) who took the MCAT for the first time in 1980. These applicants were classified using three variables: undergraduate GPA, selectivity of undergraduate college, and participation in a commercial coaching course prior to their first MCAT administration. Missing data on one or more of these variables reduced the population size to 19,564. Data on MCAT scores (first administration), undergraduate GPA, and undergraduate college attended were obtained from the files of the American Medical College Application Service (AMCAS). Selectivity of undergraduate college was determined by using the Astin index (14), which is the average SAT score of entering freshmen to each college and university. Colleges with very high selectivity were defined as those whose entering freshmen averaged 1182 or better on both parts of the SAT. High selectivity colleges were those whose freshmen averaged between 1037 and 1182 on the SAT. Medium selectivity colleges had averages between 893 and 1036, while low selectivity colleges had averages less than 893. The rationale and procedures used to group colleges in this way are described elsewhere (14). Information regarding participation in a coaching course was based, as in the first study, on a questionnaire item administered subsequent to each MCAT examination.

Within each of the 32 cell matrix formed by GPA (eight levels) and college selectivity (four levels), MCAT mean scores were computed for coached and uncoached examinees, and the differences in these mean scores were derived. Because the data were based on a population of applicants rather than a sample, statistical tests of significance were not applicable. Instead, the mean differences were examined for trends in their magnitude and pattern.

RESULTS

Table 2 shows the differences in average scores of the coached and uncoached groups by level of GPA, and Table 3 shows the differences by college selectivity. In Table 2, college selectivity was controlled by giving equal weight to the differences between groups at each of the four levels of selectivity. Similarly, in Table 3, undergraduate GPA was controlled by giving equal weight to the differences between groups at each GPA level. Each of these two tables shows an overall average difference for each MCAT subtest

TABLE 2

Differences in Average Scores of Coached and Uncoached First-Time Examinees on the Medical College Admission Test (MCAT), by Examinees' Undergraduate Grade-Point Averages (GPA), Based on Records of 19,564 Applicants to 1981–82 Entering Class*

MCAT Subtests	Undergraduate GPA†								Average Across GPA Groups
	0–2.50	2.51–2.80	2.81–3.00	3.01–3.20	3.21–3.40	3.41–3.60	3.61–3.80	3.80–4.00	
Science Knowledge									
Biology	.24	.15	.27	.58	.58	.60	.31	.42	.39
Chemistry	.32	.26	.16	.33	.39	.47	.17	.09	.28
Physics	.31	.40	.29	.57	.58	.44	.40	.43	.43
Science Problems	.23	.22	.14	.35	.35	.46	.17	.20	.26
SA: Reading	−.34	.02	−.08	−.06	−.12	.05	−.18	−.16	−.13
SA: Quantitative	.07	−.09	−.03	0	−.12	.02	−.08	−.15	−.05

* In computing these differences, equal weight was given to examinees from colleges with different levels of selectivity. A positive difference indicates a higher mean score for those enrolled in commercial coaching courses.

† For MCAT Science Knowledge and Science Problems subtests, undergraduate science GPA was used; for skills analysis subtests, undergraduate total GPA was used.

TABLE 3

Differences in Average Scores of Coached and Uncoached First-Time Examinees on the Medical College Admission Test (MCAT), by Level of Selectivity of Examinees' Undergraduate Colleges, Based on Records of 19,564 Applicants to 1981–82 Entering Class*

MCAT Subtests	Levels of College Selectivity				Average Across Selectivity
	Low	Medium	High	Very High	
Science Knowledge					
Biology	.43	.40	.48	.26	.39
Chemistry	.42	.20	.44	.04	.28
Physics	.43	.39	.59	.30	.43
Science Problems	.42	.20	.36	.07	.26
SA: Reading	−.10	−.21	−.06	−.13	−.13
SA: Quantitative	.07	−.09	−.07	−.10	−.05

* In computing these differences, equal weight was given to examinees at different levels of GPA. A positive difference indicates a higher mean score for those enrolled in commercial coaching courses.

across the undergraduate GPA and college selectivity categories.

The differential impact of commercial coaching courses on MCAT Science Knowledge and Science Problems subtests versus the Skills Analysis subtests is clear from the overall averages, and these figures support the findings of the first study. Differences on the Science Knowledge and Science Problems subtests ranged from .26 to .43 and favored coached applicants. However, the differences were trivial for the SA: Reading and SA: Quantitative subtests and tended to favor the uncoached group.

The differences by GPA level suggested that coaching courses had a somewhat smaller effect on the performance on the Science Knowledge and Science Problems subtests for applicants with low

GPAs (0–3.0) and applicants with very high GPAs (3.6–4.0). The differences by college selectivity suggested that coaching courses tended to have a smaller effect on performance on the Science Knowledge and Science Problems subtests of applicants from very selective undergraduate colleges than of those from other types of colleges.

Discussion and Interpretation

Both of the studies described here showed differences in scores of examinees on the MCAT Science Knowledge and Science Problems subtests that were at least associated with, if not due to, the examinees' taking commercial coaching courses. The study of first-time examinees showed slightly smaller differences than the study of test repeaters. Although this may be due simply to methodological differences between the studies, one might hypothesize that examinees who have taken the test and received a report of their deficiencies might use coaching courses more efficiently to remedy areas of weakness. The failure to find differences in SA: Reading subtest scores in either study and the small, unreplicated differences in SA: Quantitative subtest scores suggest that commercial coaching courses for the MCAT benefit examinees primarily by providing a structured review of the science content. Demonstration of this point requires a more focused study than the ones reported here. However, the other potential components of coaching courses—supplying correct answers, test-taking practice, maximizing motivation, controlling test anxiety, and instructing in test-taking skills—would be expected to affect all MCAT scores equally. If instruction and review are primarily responsible for the effects of coaching on MCAT scores, the courses are more appropriately described as science review programs, a term with less negative connotations.

The smaller differences in scores on the Science Knowledge and Science Problems subtests found for examinees with low scores on the SA: Reading and SA: Quantitative subtests provide serendipitous evidence of the validity of these latter two MCAT subtests. The rationale behind the development of these subtests was that the medical student and the physician, faced with a burgeoning body of scientific literature, should possess the analytical skills necessary to evaluate scientific findings critically and apply them to patient care. These same skills are, plausibly, involved in an effective review of the many and diverse science topics covered by the MCAT. They enable students preparing for the test to sift out the relevant from the irrelevant and to find and focus on the major points in each area. More data on the predictive validity of scores on the SA: Reading and SA: Quantitative subtests need to be examined. However, if these subtests are useful indicators of who will benefit from short-term MCAT review programs, they may also identify candidates who would benefit from special premedical and medical school programs, for example, science enrichment programs for minorities and others whose recruitment and retention in medical school are desired. The findings also suggest that students preparing for medical school might be aided by long-term programs that develop the skills these subtests measure.

The findings that those with very low GPAs benefited less from coaching courses in terms of their performance on the Science Knowledge and Science Problems subtests may be related to the differences by level of scores in the Skills Analysis subtests. The students' low GPAs demonstrate that they had learned